An Artificial Neural Network for Deprivation Analysis in England

Nedal Tanjaoui

BSc (Hons.) Computer Science
Honours Dissertation

Supervised by Prof. Lynne Baillie

Heriot-Watt University
School of Mathematical and Computer Sciences

September 2025

DECLARATION

I, Nedal Tanjaoui, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Nedal Tanjaoui

Date: 27.03.2025

ABSTRACT

This study investigated deprivation in England by applying machine learning techniques to predict deprivation scores at the Lower Layer Super Output Area (LSOA) level. A Multi-Layer Perceptron (MLP) model was developed using data related to the various domains of deprivation. The model achieved moderate predictive accuracy, which was expected since the selected features aligned with the themes of the Index of Multiple Deprivation (IMD) but did not exactly match its official indicators.

SHapley Additive Explanations (SHAP) analysis showed that employment, income, and education related features were the strongest predictors. However, some features, despite being official IMD indicators, had a weaker than expected influence. This suggested that broader conditions played a more significant role in shaping deprivation.

Geographical analysis revealed that urban areas experienced greater reductions in deprivation scores than rural areas, supporting research on urban investment and regeneration. An unexpected finding was the limited impact of crime-related features on predicted deprivation scores, despite crime being a key IMD domain. This could suggest that crime rates were closely linked to other socioeconomic factors, which may have reduced their individual influence on predictions.

While machine learning provided valuable insights, the study highlighted its limitations and recommendation for further work. Overall, this study contributed to the discussion on deprivation, emphasizing the benefits of policy interventions using machine learning driven approaches.

ACKNOWLEDGEMENTS

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Motivation

The disparity in deprivation between areas is a deep-rooted issue in England that affects many aspects of society, and there has been a vast amount of research that concludes that socioeconomic background is a key determinant of opportunity, success and access to key services such as health and education. The results of one such study that investigated the impact of deprivation on health suggested that England's poorest areas face a 'double burden of inequality,' with both reduced life expectancies and greater unpredictability in the timing of death [Mayhew et al. 2020]. When students apply for university courses, or candidates apply for jobs, they may be asked for information about their background so their circumstances can be evaluated and the barrier to entry adjusted. There are many other systems in place in attempt to 'level the playing field' such as diversity initiatives targeting under-represented groups.

While these measures aid in mitigating the disadvantages that can arise from applicants coming from deprived communities, more effort needs to be focused on understanding and tackling the problems at the source. Understanding the characteristics of an area that makes it more deprived than others can help inspire government policy and strategy changes that could help create an environment where individuals from all backgrounds have a fair chance to succeed.

There are already existing methods that measure a location's deprivation relative to others, but these approaches require vast amounts of resources, domain knowledge and expertise. The process of manually selecting statistics and tuning their weightings is time consuming or even impractical if the number of relevant data attributes is large enough. Despite updated statistics from the 2021 census [Office for National Statistics (ONS) 2021], the most recent release of the English indices of multiple deprivation was in 2019 that still used 2011 census data [Office for National Statistics (ONS) 2011] for some indicators. Implementing a machine learning approach for deprivation indexing is an opportunity to quicken the process of reporting key findings about the new data and changes in relation to previous releases. However, reducing the reliance on intensive domain expertise makes the potential for a powerful tool to complement existing methods rather than a pure replacement because the model is trained on pre-existing calculations.

1.2 Aim and Objectives

Objective 1: Develop a Neural Network to Compute the Deprivation Score Using 2011 Census Data.

Designing and implementing a NN using 2011 census data and the pre-calculated deprivation

scores for geographic locations across England.

Objective 2: Evaluate the Performance of the Model in Computing the Deprivation Score

Applying evaluation methods and metrics to analyze the accuracy of the predictions and guide refinements.

Objective 3: Apply the Model to the 2021 Census Data and Lower Layer Super Output Areas (LSOAs) to Analyze Changes

Applying the model to the 2021 Census data and LSOAs to produce updated deprivation indices that will allow for a comparative analysis between 2011 and 2021.

Objective 4: Simulate and Assess the Impact of Hypothetical Policy Changes

Simulate hypothetical scenarios where certain input features will be modified to mimic potential policy changes.

1.3 Organisation

After stating motivations and objectives in this chapter, I will provide a comprehensive review of existing literature and background information to set the context for my work. Then in the third chapter I will detail the project requirements and methodology conducted. In the fourth chapter I will present the results of the methodology, which are analysed in the fifth chapter. The sixth chapter delves into the broader implications on my analysis and outlines the limitations of this work, as well as recommendation for further research. Chapter seven concludes my work, revisiting the objectives. An ethical discussion, and a GitHub link to the code and datasets used, are provided in the appendices.

2 BACKGROUND

In this chapter I will provide an overview of the concept, data sources and methodologies that are central to my project. I will begin by introducing Lower Layer Super Output Areas (LSOA's) in section 2.1. Then in section 2.2 I will delve into the Index of Multiple Deprivation, outlining its purpose and methods of assessing relative deprivation across LSOAs, as well as its limitations. Then I will introduce the Neural Network that I plan to implement, the Multi-Layer Perceptron, and a method to interpret the results in section 2.5. Finally, I will review related work.

2.1 Lower Layer Super Output Areas (LSOAs)

The concept of Output Areas (OAs) was first introduced following the UK's 2001 Census [Office for National Statistics (ONS) 2001]. In the context of Census geography [Office for National Statistics (ONS) 2024], OAs are the smallest unit for geographical area covering roughly 125 households (usually between 40 and 250) and a population of 300 (usually between 100 and 625). To handle population changes in successive censuses, OA's that still met the previously mentioned thresholds remained unchanged and the rest were either merged or split if the population and number of households fell outside certain thresholds. LSOAs are the next smallest unit, usually made up of 5 OAs and therefore covering a population of around 1500 and roughly 625 households. As of the 2021 Census, there are 33,755 LSOA's in England.

2.2 Index of Multiple Deprivation (IMD)

Townsend [1979] defines people being deprived as if they 'lack the types of diet, clothing, housing, household facilities and fuel and environmental, educational, working and social conditions, activities and facilities which are customary'. Townsends account and the idea that deprivation comes in multiple domains inspired the methods used to form the Index of Multiple Deprivation. The IMD is the measure of relative deprivation at the LSOA level. It is designed to provide a comprehensive outlook on an LSAO's deprivation in comparison to others by combining and weighing 7 different domain scores (shown in figure 1) to produce a deprivation score that is then used to rank each LSOA from 1 (most deprived) to 32,844 (least deprived). The IMD is widely used by government agencies, researchers, and policymakers to identify areas that may require additional resources and policy initiatives [Deas et al. 2003]. It helps the allocation of funding, the shaping of policies and the evaluation of the effectiveness of the different types of support an area receives.

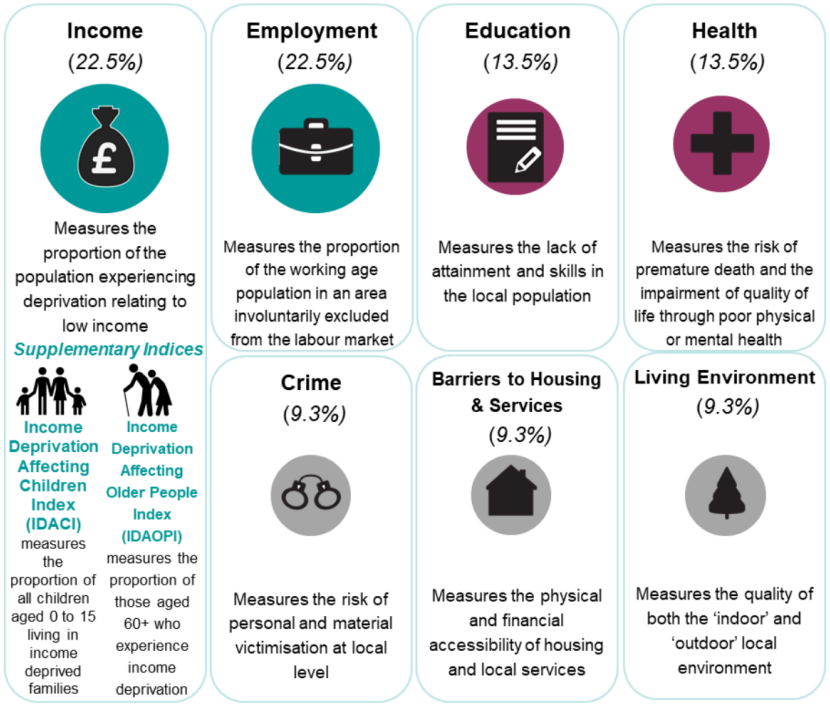

Fig. 1. The 7 domains of the Index of Multiple Deprivation (IMD) and their weightings [Ministry of Housing, Communities and Local Government 2019]

With deprivation being a multi-dimensional concept and experienced in multiple domains [Abascal et al. 2022], a great amount of thought has been taken to select the most appropriate indicators and statistics to form the score of each domain. I will now review the English Indices of Deprivation 2015 technical report [Ministry of Housing, Communities and Local Government 2015] to identify features and indicators that were used to produce each domain score. Although many of the statistics related to the indicators aren't publicly available, understanding the reasoning and themes behind the selected indicators will help inspire my feature selection process, aiding in the identification of relevant features that align with these domains in the 2011 Census dataset.

2.2.1 Income. Because some components of deprivation stem from low income [Townsend 1987], theoretically speaking, income should not be a domain of the IMD. However, since there is a lack of research and data at the LSOA level on indicators related to some of the items mentioned in Townsends definition, the income domain is a very reasonable substitute. The Income Deprivation Domain (IDD) score is calculated using the proportion of people in the LSOA that experience low-income deprivation which is defined as 'people that are out-of-work, and those that are in work but who have low earnings' [Ministry of Housing, Communities and Local Government 2015]. The indicators for this are whether adults and

children are receiving forms of means tested income support and benefits such as Working Tax Credit, Child Tax Credit and Pension Credit.

2.2.2 Employment. As is the case with the IDD, the Employment Deprivation Domain (EDD) measures the proportion of working aged people (in an LSOA) that experience employment deprivation. The EID 2015 technical report defines this as 'people who would like to work but are unable to do so due to unemployment, sickness or disability, or caring responsibilities'. The indicators for the EDD include the receipt of income support and benefits related to those terms such as Employment and Support Allowance, Jobseeker's Allowance and Carer's Allowance. However, unlike the IDD, those that don't meet the age criteria are excluded (18 to 59 for Women, and 18 to 64 for men).

2.2.3 Education, Skills and Training (EST). The EST Deprivation Domain (ESTDD) score reflects the 'lack of attainment and skills in the local population' [Ministry of Housing, Communities and Local Government 2015]. The ESTDD is split into two sub-domains: adults and young people with separate indicators for each, although they follow the same themes. Indicators for young people focus on test scores and the proportion entering higher education, while indicators for adults pertain to qualifications and English proficiency.

2.2.4 Health. The Health Deprivation and Disability Domain (HDD) assesses the risk of premature death and the impairment of quality of life due to poor physical or mental health. It focuses on measuring current morbidity, disability, and premature mortality but excludes factors related to behavior or environment that might predict future health outcomes. The indicators are related to mental and physical health, premature death and emergency hospital admission rates. Statistics that are relevant include the proportion of the population that receive forms of disability or health benefits.

2.2.5 Crime. The previous domains primarily focus on deprivation at the household and personal level, making them significant, as reflected in their weightings. However, area-level domains, such as crime, also significantly influence households' exposure to deprivation [Abascal et al. 2022]. In the context of the English IMD, the Crime Domain assesses the 'risk of personal and material victimization within a local area' [Ministry of Housing, Communities and Local Government 2015]. It is based on indicators that capture the occurrence rates per 1,000 units: violence, theft, and criminal damage are measured per 1,000 people, while burglary is measured per 1,000 properties. These indicators provide a comprehensive view of how crime impacts community-level deprivation.

2.2.6 Barriers to Housing and Services. The Barriers to Housing and Services Domain assesses the accessibility of housing and essential services. It is divided into two subdomains: Geographical Barriers and Wider Barriers. Geographical Barriers focuses on the proximity of key services, measuring the average road distance to the nearest amenities such as post offices, schools, supermarkets, and GPs for residents within an LSOA. The Wider Barriers domain

addresses housing-related deprivation, including household overcrowding, homelessness rates based on local authority data, and housing affordability. Together, these subdomains provide a comprehensive view of the challenges individuals and households face in accessing housing and the necessary services.

2.2.7 Living Environment. Living Environment Deprivation Domain evaluates the quality of the local environment through two subdomains: Indoors and Outdoors. The Indoors subdomain relates to deprivation at the household level focusing on housing quality by measuring the proportion of houses without central heating and the percentage of social and private homes that fail to meet the Decent Homes standard. In contrast, the Outdoors subdomain focuses on area-level deprivation assessing environmental factors, including air quality and the rate of road traffic accidents involving injury to pedestrians and cyclists. Together, these indicators provide a comprehensive measure of how both housing conditions and external environmental factors contribute to deprivation.

2.2.8 Limitation of the IMD. Deas et al. [2003] critique the methods used to form the IMD, highlighting several limitations that still exist in the 2015 IMD. The limitation that is argued to be the most obvious is the issue of double counting of certain indicators across multiple domains. For example, claimants of Employment and Support Allowance, Incapacity Benefit and Severe Disablement Allowance is used as an indicator for both the Employment and Health Deprivation Domains. The IMD technical report defends this by arguing that it's 'desirable and appropriate to measure situations where deprivation occurs in more than one dimension' [Ministry of Housing, Communities and Local Government 2015]. However, this fails to invalidate the criticism because scores from multiple domains are aggregated into a single IMD score, so using the same indicators across multiple domains inflates the IMD for areas where these benefits have a higher occurrence. This reduces the accuracy in identifying needs and could result in a disproportionate allocation of resources to those areas.

A second concern is the reliance on data about the take-up of benefits. The proportion of entitled non-recipients of benefits can vary by geographical area, so ideally you would adjust the data to take that into account. However, for the DWP's income related benefits, there are no estimates for how the take-up of benefits varies between areas, so no geographical adjustments can be made [Department for Work and Pensions 2012]. Again, this reduces the validity of the IMD in terms of evaluating an LSOA's relative deprivation.

Finally, as previously mentioned, the IMD requires extensive time, expertise, and resources for its calculation. This results in high costs and long time frames, which is reflected by the infrequency of updates to the IMD. A recent study by the Social Metric Commission found an increase of 2.1 million people in poverty since 2019/20 [Social Metrics Commission 2024], but with the government yet to produce an update on the English Indices of Deprivation, a comprehensive understanding of how deprivation patterns have shifted in response to recent

socioeconomic changes since the pandemic remains unavailable. This gap highlights the need for automated methods to measure deprivation to update the IMD more frequently.

2.3 Census Dataset

The Census is a nationwide survey conducted every ten years in England and Wales to collect information about the population and its characteristics [Office for National Statistics (ONS) 2011]. The Census covers a wide range of topics, including demographics, housing, employment, health, and education. For this project, the 2011 Census will provide the data required to be able to select features that align with the 7 domains of the IMD.

2.4 Multi-Layer Perceptron (MLP)

An Artificial Neural Network (ANN) is a machine learning model inspired by the structure of the human brain. An MLP is a type of ANN, consisting of an input layer, then one or more hidden layers, and finally an output layer, as shown in figure 2.

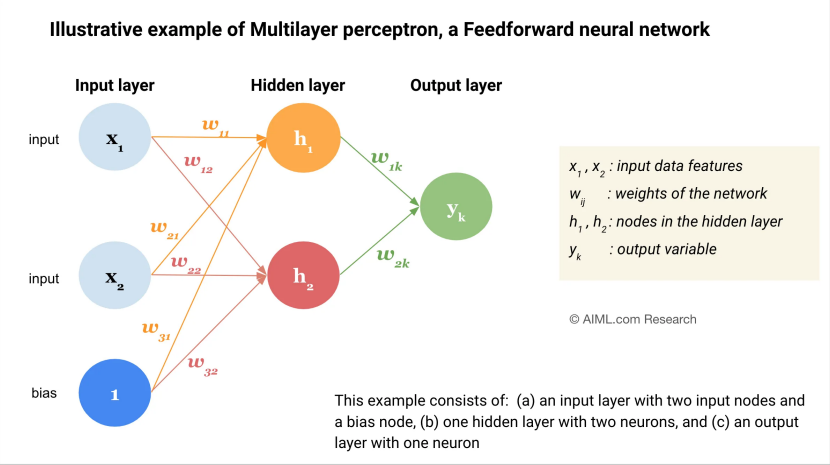

Fig. 2. Multi-Layer Perceptron [AIML 2024]

Each neuron in the input layer (xi) corresponds to a feature (or attribute) of the input data. There are no computations performed at this layer, and data is simply passed to the hidden layer. The hidden layer is where calculations on this data are performed. Each neuron in this layer receives an input from all the neurons in the previous layer, multiplying each by a weight (wi), adding biases (b), then applying an activation function (f(z)) to the sum, as illustrated in figure 3.

Neurons in subsequent hidden or output layers perform the same calculations, using the

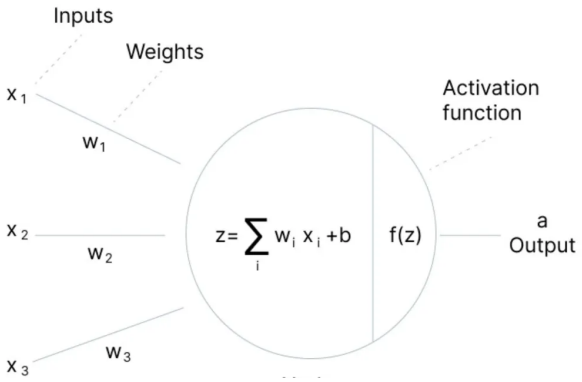

Fig. 3. Calculations performed by neurons in the hidden and output layers [AIML 2024]

previous layer's outputs, this technique is called forward propagation. Then the output layer produces a prediction that is compared to the actual target value to quantify the difference, using that information to adjust the weights and biases (backpropagation). This process is repeated for multiple iterations until a satisfactory level of accuracy is reached.

2.4.1 Benefits of this approach. MLP algorithmically learns the relationships between features and the deprivation index, eliminating the need for manual weighting and feature (indicator) selection that the existing method for calculating the IMD requires. This means that this approach can be used to produce more frequent updates of the IMD as new data is released to more accurately reflect current socio-economic conditions.

2.4.2 Limitations. Using a neural network requires existing calculations of the IMD to train the model, meaning that this approach complements rather than replaces existing methods. Also, due to the automated approach, policy makers and stakeholders may have less trust in recommendations made using this model.

2.5 SHapley Additive exPlanations (SHAP)

SHAP is a method used to interpret machine learning models by assigning importance scores to input features, developed by Lundberg and Lee [Lundberg and Lee 2017]. It is based on Shapley values, a concept from cooperative game theory developed by Lloyd Shapley (1953). In the context of machine learning, features are treated as "players", and their contributions to the model's output are quantified using Shapley values.

The SHAP value of a feature i is the average marginal contribution of that feature across all possible feature subsets:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right]$$

Where F is the set of all features, S is a subset of features excluding i, $f(S)$ is the model's output when only the features in S are considered, ϕ_i represents the SHAP value for feature i.

2.5.1 Applications for this work. SHAP can be used to gain insights about the most impactful features on the MLP's predictions. SHAP also identifies the relationship between feature values and model predictions. For example, SHAP visualisations can show whether lower education attainment had a positive or negative impact on predicted IMD scores.

2.6 Related Work

In this section I will review previous work related to my objectives. Specifically, the application of machine learning techniques as a more effective approach to traditional statistical methods.

2.6.1 Logistic Regression Versus Machine Learning Approaches. Montebruno et al. [2020] set out to classify entrepreneurial status (either worker or entrepreneur) for individuals that didn't report it in British historical census data from 1851 to 1881, available from the Integrated Census Microdata [Schurer et al. 2024]. Logistic regression and various machine learning algorithms were trained and tested on later editions of the census where classification is known.

Model accuracy was calculated using: $\frac{\text{True Positives+True Negatives}}{\text{Total Predictions}}$

Logistic regression, the base-line method, achieved an accuracy of 0.82, making it the worst performing model. The ANNs, specifically the Sequential Neural Network and Recurrent Neural Network (with a Long Short-term Memory layer) performed the best with respective accuracies of 0.996 and 0.998 (3 d.p). This proves that ML, and especially DL, can offer substantial improvements over traditional approaches. Also, they found that integration of text-based features into ML models, such as bag-of-words and text embeddings, contributed significantly to accuracy. [Montebruno et al. 2020] recommend that researchers should test multiple ML models and features before settling on a particular approach.

2.6.2 Multi-Layer Perceptron Approaches. De Fausti et al. [2022] used a Multi-Layer Perceptron model to predict the attained level of education (ALE) for Italian residents. The actual ALE is only known for about 5% of the population though sample surveys. For the remaining 95% the ALE is estimated using traditional statistical methods (Log-Linear models) by combing administrative records, data from the 2011 census, and information from the surveys. Like the IMD, this 'requires an expensive initial phase of data analysis and treatment to achieve an accurate prediction' [De Fausti et al. 2022].

With the objective of making the process of predicting the ALE more efficient and effective, this study implemented an MLP model to conduct two different experiments. In the first experiment, the MLP was applied within the same informational framework as the official

method, using the same preprocessed covariates and aggregation levels with the goal of improving the quality of predictions. In the second experiment, the MLP was trained on raw, unaggregated data, relying on the model's network's ability to identify patterns and relationships. The objective of the second experiment was to purely assess whether the implementation was a valid approach to automate and quicken the ALE estimation process.

Results showed that there was no significant difference in predictive performance between the existing method and the MLP model. However, in the case of the second experiment where raw data was used, the model produced comparable predictive performance, therefore achieving the objective of being a more streamlined approach of estimating the ALE.

2.7 Conclusion

This chapter has provided an overview of the key concepts, data sources, and methodologies related to my research. I highlighted the complexity in calculating the IMD and its reliance on time-consuming manual weighting and feature selection. I've discussed how machine learning techniques have previously shown promise in addressing these limitations by reviewing work by Montebruno et al. [2020] and De Fausti et al. [2022]. It is clear that Machine learning models, specifically MLPs, used alongside the existing method of calculating the IMD, could enable more frequent updates or even real-time analysis of deprivation levels as updated statistics are released, allowing policy makers to identify locations in need of resources quicker.

3 METHOD

This section outlines the requirements for this project, and the methods employed to execute them. The process begins with the selection of the most relevant features from the Census and crime datasets. Next, data preprocessing techniques are applied to prepare the data for training. Then, an MLP model is implemented and its parameters refined using results from evaluation metrics. The model is then applied to the 2021 Census and updated crime datasets, and the methods used to analyse the results are discussed.

3.1 Project Requirements Analysis

The project aims to train an MLP model to predict IMD scores across LSOAs in England. This section outlines the project requirements.

3.1.1 Data Requirements. Data sources that need to be used include the Census Data from 2011 and 2021 as well as crime data at the LSOA level, provided by the Office for National Statistics (ONS) and police forces. The data needs to align with the different domains of deprivation, or ideally match the indicators that were used to calculate the IMD scores. Additionally, data regarding LSOA boundaries is needed for creating geographic visualizations.

3.1.2 Model Requirements. The neural network model design needs to accommodate the following: The model will be trained on input features derived from census data as well as crime data. A Multi-Layer Perceptron (MLP) architecture, with an input layer, a number of hidden layers that balances capacity with compuational efficiency, and an output layer. The Rectified Linear Unit (ReLU) will be used to overcome the vanishing gradient problem and provide better training performance for deep networks. The Adaptive Moment Estimation (Adam) optimizer will be used for efficient training, with Mean Squared Error (MSE) as the loss function to suit the regression task.

3.1.3 Geographic Visualization Requirements. Visualizing deprivation levels geographically is critical to understanding spatial patterns. To display IMD scores across England's LSOAs, maps will be generated. The map must allow for panning and zooming at the LSOA level. GeoJSON files will be used to plot the LSOA boundaries.

3.1.4 Model Interpretation Requirements. SHapley Additive exPlanations (SHAP) will be utilized to interpret the model's decisions by assigning importance to each feature. A Beeswarm plot will be used to visualize this. A subset of training data (1,000 samples) will be used to compute SHAP values to allow for reasonable run times.

3.1.5 Justifications for changes to the Project's Aims and Objectives. Objective 4, which was to simulate hypothetical policy changes by modifying feature values, was initially included because it allows for the analysis of how feature values impact IMD predictions. This is replaced with the SHAP model explainer, which achieves the same goal more efficiently.

The original objectives don't enable a spatial analysis of the distribution of IMD scores, so the geographic visualization requirement was added.

The Census dataset doesn't contain any crime related data, but police forces have made crime statistics at the LSOA level publicly available. This data will be used to complement the Census statistics for model training.

3.2 Feature Selection

The feature selection process was carried out to identify the features that align or ideally match the indicators mentioned in the IMD 2015 technical report. From the Census 2011 dataset and crime statistics, several relevant features identified.

3.3 Data Pre-processing

The preprocessing phase involved multiple steps to ensure the dataset was structured appropriately for training.

3.3.1 Data Loading. The datasets containing broadly relevant features were loaded using the Pandas library. Then, the columns were merged into a single table using LSOA codes as the index. A combined table was produced containing 32844 instances (LSOAs).

3.3.2 Feature Engineering. Similar to the methods mentioned in the IMD 2015 technical report, values were converted into percentages, with the denominator being either the number of households or total population in the LSOA. However, in the case of features related to crime rates, occurrences per 1000 people was calculated for the 2013/14 year because that was the approach used for those indicators for the IMD. Some features were derived from multiple columns. For example [4], the proportion of overcrowded houses was calculated by adding together the number of homes with an occupancy rating of -1 to those with -2 or less, then dividing the sum by the total number of households to get a percentage. This was again the same method used to obtain the "household overcrowding" indicator mentioned in the IMD 2015 technical report. A total of 16 features were derived, as presented alongside their corresponding IMD domain in table 1. The final table contained 17 columns (including the IMD score) and all 32844 instances with no null values.

```
occupancy_rooms["proportion of overcrowed houses"] = (occupancy_rooms.iloc[:, 8] + occupancy_rooms.iloc[:, 9]) / occupancy_rooms.iloc[:, 4]
```

```
occupancy_rooms.head()
```

	date	geography	geography code	Rural Urban	Occupancy Rating: All categories: Occupancy rating (rooms); measures: Value	Occupancy Rating: Occupancy rating (rooms) of +2 or more; measures: Value	Occupancy Rating: Occupancy rating (rooms) of +1; measures: Value	Occupancy Rating: Occupancy rating (rooms) of 0; measures: Value	Occupancy Rating: Occupancy rating (rooms) of -1; measures: Value	Occupancy Rating: Occupancy rating (rooms) of -2 or less; measures: Value	proportion of overcrowed houses
0	2011	Darlington 001B	E01012334	Total	962	804	103	43	12	0	0.012474
1	2011	Darlington 001C	E01012335	Total	620	459	103	48	10	0	0.016129
2	2011	Darlington 001D	E01012366	Total	860	661	120	60	18	1	0.022093
3	2011	Darlington 001E	E01033481	Total	554	329	146	62	17	0	0.030686
4	2011	Darlington 001F	E01033482	Total	650	431	162	47	6	4	0.015385

Fig. 4. Occupancy rating calculation

Feature Description	IMD Domain
Proportion of overcrowded houses	Barriers to Housing and Services
Proportion of households with no central heating	Living Environment Deprivation
Proportion of people who cannot speak English well or at all	EST
Proportion of people with Level 2 qualifications or below	EST
Proportion of people with Level 3 qualifications or below	EST
Proportion of households with no adults in employment	Employment Deprivation
Proportion of households with one person with a long-term health problem or disability	Employment Deprivation
Proportion of people providing 20 or more hours of unpaid care	Employment Deprivation
Proportion of people in bad or very bad health	Health Deprivation
Proportion of people whose day-to-day activities are limited a lot	Health Deprivation
Proportion of people working 15 hours or less per week	Income Deprivation
Proportion of unemployed individuals	Income Deprivation
Burglary rate per 1000 people	Crime
Violence rate per 1000 people	Crime
Criminal damage and arson rate per 1000 people	Crime
Theft rate per 1000 people	Crime

Table 1. Feature descriptions and their corresponding IMD domains

3.3.3 Standardisation of Data. The final preprocessing step was the standardization of data. Due to varying feature scales, the data needs to be standardized so that all features have a mean of 0 and standard deviation of 1. This allows features to contribute equally in the learning process of the Multi-Layer Perceptron model. Standardization was chosen instead of normalisation because it reduces the impact of extreme values and distributes the data closer to normal which should improve model performance. Due to the dataset containing many instances, a train-test split of 75/25 was chosen. The StandardScaler and train_test_split libraries were used as shown in figure 5.

```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

X = merged.drop('Index of Multiple Deprivation (IMD) Score', axis=1)
y = merged['Index of Multiple Deprivation (IMD) Score']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)

print(f"Training data: {len(X_train)}")
print(f"Testing data: {len(X_test)}")

scaler_X = StandardScaler()
scaler_y = StandardScaler()

X_train_scaled = scaler_X.fit_transform(X_train)
X_test_scaled = scaler_X.transform(X_test)

y_train_scaled = scaler_y.fit_transform(y_train.values.reshape(-1, 1))
y_test_scaled = scaler_y.transform(y_test.values.reshape(-1, 1))
```

```
Training data: 24633
Testing data: 8211
```

Fig. 5. Standardization of data

3.4 Multi-Layer Perceptron Design

The Multi-Layer Perceptron (MLP) model implemented [6] was designed using the PyTorch machine learning library to predict the Index of Multiple Deprivation (IMD) Score based on the 16 feature input.

3.4.1 Model Architecture. The model contains an input layer of size 16, then two hidden layers each with 64 neurons, then an output layer. This provides sufficient capacity to learn pattern while also considering computational efficiency. The Rectified Linear Unit (ReLU) was chosen as the activation function for the hidden layers because it mitigates the vanishing gradient problem.

3.4.2 Model Hyper-Parameters. The Adaptive Moment Estimation (Adam) optimizer was used as it is commonly viewed as the most performant optimiser in recent times. Because predicting the IMD score is a regression task, the loss function used was Mean Squared Error (MSE). 1,000 epochs and a batch size of 64 were chosen to maintain a balance between model effectiveness and computational efficiency.

```python
import torch
import torch.nn as nn
import torch.optim as optim

gpu = torch.device("cuda" if torch.cuda.is_available() else "cpu")

X_train_tensor = torch.tensor(X_train_scaled, dtype=torch.float32).to(gpu)
X_test_tensor = torch.tensor(X_test_scaled, dtype=torch.float32).to(gpu)
y_train_tensor = torch.tensor(y_train_scaled, dtype=torch.float32).to(gpu)
y_test_tensor = torch.tensor(y_test_scaled, dtype=torch.float32).to(gpu)

class MLPModel(nn.Module):
    def __init__(self, inputs, hidden, output):
        super(MLPModel, self).__init__()
        self.hidden_layer1 = nn.Linear(inputs, hidden)
        self.relu = nn.ReLU()
        self.hidden_layer2 = nn.Linear(hidden, hidden)
        self.output_layer = nn.Linear(hidden, output)

    def forward(self, x):
        x = self.relu(self.hidden_layer1(x))
        x = self.relu(self.hidden_layer2(x))
        x = self.output_layer(x)
        return x

inputs = X_train_tensor.shape[1]
hidden = 64
output = 1

model = MLPModel(inputs, hidden, output).to(gpu)

criterion = nn.MSELoss()
optimizer = optim.Adam(model.parameters())

epochs = 1000
batch_size = 64

for epoch in range(epochs):
    for i in range(0, len(X_train_tensor), batch_size):
        X_batch = X_train_tensor[i:i+batch_size]
        y_batch = y_train_tensor[i:i+batch_size]

        optimizer.zero_grad()
        predictions = model(X_batch)
        loss = criterion(predictions, y_batch.view(-1, 1))
        loss.backward()
        optimizer.step()

    if (epoch + 1) % 100 == 0:
        print("Epoch", epoch + 1, "/", epochs, "Loss:", round(loss.item(), 4))
```

Fig. 6. MLP Design

3.5 Application of Model to Updated Data

To apply the trained model to the 2021 Census statistics, the same preprocessing steps were taken with the exact same features to ensure consistency, creating a new dataframe containing 2021 LSOA's and their predicted IMD scores [7]. The updated crime statistics were from the 2022/23 year to mitigate potential biases caused by lockdowns.

```python
lsoa_column2021 = merged2["LSOA"]
merged2 = merged2.drop(merged2.columns[[0, 13]], axis=1)

X_merged2 = merged2.copy()

# Standardize features
X_merged2_scaled = scaler_X.transform(X_merged2)

# Convert to PyTorch Tensor
X_merged2_tensor = torch.tensor(X_merged2_scaled, dtype=torch.float32).to(gpu)

# Generate Predictions
model.eval()
with torch.no_grad():
    predictions = model(X_merged2_tensor)

# Convert Predictions Back to Original Scale
predictions_np = scaler_y.inverse_transform(predictions.cpu().numpy())

# Create results DataFrame
results_df2021 = pd.DataFrame({
    'LSOA': lsoa_column2021,
    'Predicted IMD Score': predictions_np.flatten()
})
```

Fig. 7. Application of MLP to 2021 Census statistics

3.6 SHAP Visualizations

To gain insight into the most influential features of the MLP, SHAP (SHapley Additive exPlanations) was used to interpret the Multi-Layer Perceptron Model (MLP) [8]. The SHAP explainer computes feature importance values and visualizes their impact on model predictions. To consider computational efficiency, a subset of 1000 samples from the training data was used to initialize a shap.GradientExplainer. The SHAP values were computed and a Beeswarm plot was produced.

```python
X_test_sample = X_test_tensor[:1000].cpu().numpy()

explainer = shap.GradientExplainer(model, X_train_tensor[:1000])

shap_values = explainer.shap_values(X_test_tensor[:1000])
shap_values = shap_values.squeeze(-1)
shap.summary_plot(shap_values, X_test_sample, feature_names=X.columns)
```

Fig. 8. SHAP gradient explainer

3.7 Geographic Visualizations

To visualize the geographic distribution of deprivation levels throughout England, multiple choropleth maps were generated using the Folium library. Initially, a static choropleth plot using the Matplotlib library was considered; however, Folium offers several advantages. Folium

allows for interactive zooming and panning, making it much more user-friendly than static Matplotlib maps. This is useful for analysing large geographical areas with relatively small geographical units. Folium maps can also be exported to HTML, making it possible to share and embed in websites with the full interactivity features.

A total of four different maps were produced for analysis, the first being a map of the 2011 LSOA boundaries with the actual 2015 IMD scores. A CSV (comma separated values) file containing IMD 2015 scores and LSOA codes was merged with a GeoJSON file (loaded using the GeoPandas library) that contains 2011 LSOAs boundaries and their polygon coordinates [9]. The merged data was used to generate the Folium choropleth map using the parameters shown in figure [10].

```python
import geopandas as gpd
import matplotlib.pyplot as plt

# Load LSOA boundary GeoJSON
lsoa_map = gpd.read_file("lsoa.geojson")  # Update path if needed

# Filter out Welsh LSOAs
lsoa_map_england = lsoa_map[lsoa_map['LSOA11CD'].str.startswith('E')]

# Rename LSOA ID column to match
scores_imd.rename(columns={'LSOA code (2011)': 'LSOA11CD'}, inplace=True)

# Merge boundaries with deprivation scores
lsoa_map_england = lsoa_map.merge(scores_imd, on="LSOA11CD", how="left")
```

Fig. 9. Merging 2015 IMD scores with the GeoJSON file

```python
import folium

# centre on England
m = folium.Map(location=[52.3555, -1.1743], zoom_start=6)

folium.Choropleth(
    geo_data=lsoa_map_england,
    data=scores_imd,
    columns=["LSOA11CD", "Index of Multiple Deprivation (IMD) Score"],
    key_on="feature.properties.LSOA11CD",
    fill_color="OrRd",
    fill_opacity=0.7,
    line_opacity=0.2,
    legend_name="Index of Multiple Deprivation (IMD) Score"
).add_to(m)

# Save to HTML file
m.save("england_deprivation_map.html")
m
```

Fig. 10. Choropleth map for actual 2015 IMD scores of 2011 LSOAs

A second map was generated to visualize the predicted 2015 IMD scores, I applied my neural network model to the entire dataset [11]. First, the dataset was standardized with the same scaler used during training, then predictions were generated and converted back to original scale and a new CSV file was created to be able to repeat the process of producing a Folium map [12].

```
# Evaluation mode
model.eval()

# Standardize dataset
X_full = merged.drop(columns=['Index of Multiple Deprivation (IMD) Score'])
X_full_scaled = scaler_X.transform(X_full)

# Convert to PyTorch tensor and move to GPU
X_full_tensor = torch.tensor(X_full_scaled, dtype=torch.float32).to(gpu)

# Make prediction using trained model
with torch.no_grad():
    y_pred_scaled = model(X_full_tensor)

# Convert predictions back to original scale
y_pred = scaler_y.inverse_transform(y_pred_scaled.cpu().numpy())

# Create DataFrame with LSOA and Predicted IMD Score
results_df = pd.DataFrame({
    'LSOA': lsoa_column,
    'Predicted IMD Score': y_pred.flatten()
})

# Save to CSV
results_df.to_csv('predicted_imd_scores.csv', index=False)
```

Fig. 11. Applying MLP to the full dataset

```
import folium
from folium.plugins import Fullscreen
# Load IMD 2011 predictions
results_df = pd.read_csv("predicted_imd_scores.csv")

# Load LSOA GeoJSON file
gdf = gpd.read_file("lsoa.geojson")

# Match column name
gdf = gdf.rename(columns={'LSOA11CD': 'LSOA'})
merged_gdf = gdf.merge(results_df, on="LSOA", how="left")

# Centre of england
m = folium.Map(location=[52.3555, -1.1743], zoom_start=6, tiles="cartodbpositron")

folium.Choropleth(
    geo_data=merged_gdf,
    name="IMD Score",
    data=merged_gdf,
    columns=["LSOA", "Predicted IMD Score"],
    key_on="feature.properties.LSOA",
    fill_color="YlOrRd",
    fill_opacity=0.7,
    line_opacity=0.2,
    legend_name="Predicted IMD Score"
).add_to(m)

Fullscreen().add_to(m)

m.save("imd_map.html")
m
```

Fig. 12. Creating a map to visualise IMD score predictions for 2011 Census and 2013/14 crime data

A third map was made to visualise the results of the predictions made using the 2021 Census and 2022/23 crime statistics. A GeoJSON file for 2021 LSOA boundaries was unavailable so a shapefile was used instead, though for this use case there is no effect on the visualisation [13].

```
# Load LSOA Shapefile
gdf2021 = gpd.read_file("imd2021/lsoa2021.zip")

# match column names
gdf2021 = gdf.rename(columns={'LSOA11CD': 'LSOA'})
merged_gdf2021 = gdf.merge(results_df2021, on="LSOA", how="left")

# Centre
m = folium.Map(location=[52.3555, -1.1743], zoom_start=6, tiles="cartodbpositron")

folium.Choropleth(
    geo_data=merged_gdf2021,
    name="IMD Score",
    data=merged_gdf2021,
    columns=["LSOA", "Predicted IMD Score"],
    key_on="feature.properties.LSOA",
    fill_color="YlOrRd",
    fill_opacity=0.7,
    line_opacity=0.2,
    legend_name="Predicted IMD Score"
).add_to(m)

Fullscreen().add_to(m)

m.save("imd_map.html")
m
```

Fig. 13. Creating a map to visualise IMD score predictions for 2021 Census data

To analyse the changes in IMD scores between the two datasets, the difference in IMD score for each LSOA was calculated and visualised with a map [14]. This allowed for a direct comparison of deprivation levels over time. However, as explained in the background chapter, LSOA boundaries change after each Census, so only the LSOAs common between the datasets could be plotted. The 2011 LSOA boundaries were used for the visualisation with approximately 1000 LSOAs missing.

```
# Calculate the change in IMD scores
merged_results_df['IMD_change'] = merged_results_df['Predicted IMD Score_y'] - merged_results_df['Predicted IMD Score_x']

# Merge the GeoDataFrame with the merged_results_df based on LSOA
lsoa_map_england = lsoa_map_england.rename(columns={'LSOA11CD': 'LSOA'})
merged_change_gdf = lsoa_map_england.merge(merged_results_df, on="LSOA", how="left")

# centre map
m = folium.Map(location=[52.3555, -1.1743], zoom_start=6, tiles="cartodbpositron")

folium.Choropleth(
    geo_data=merged_change_gdf,
    name="IMD Change",
    data=merged_change_gdf,
    columns=["LSOA", "IMD_change"],
    key_on="feature.properties.LSOA",
    fill_color="RdYlBu_r",
    fill_opacity=0.7,
    line_opacity=0.2,
    legend_name="Change in IMD Score (2021 - 2011)",
).add_to(m)

Fullscreen().add_to(m)

m.save("imd_change_map.html")

m
```

Fig. 14. Creating a map to analyse the changes in IMD scores between the two datasets

3.8 Summary

This chapter detailed the implementation of the MLP model, including feature selection, data preprocessing, and model architecture. The IMD scores were predicted for 2011 and 2021 Census and crime data, with SHAP visualizations to provide insights into key drivers of deprivation. Interactive Folium maps were generated to visualize the distributions of deprivation, compare predicted scores with actual 2015 IMD values, and analyse changes over time. In the next chapter, the results of the implementations are presented.

4 RESULTS

This chapter presents the results of the methodology described in the previous chapter. First, the accuracy of the Multi-Layer Perceptron (MLP) model is evaluated by comparing predicted IMD scores to actual values from the 2011 dataset. This is followed by visualisations of SHAP values, and finally, geographical visualizations of the distribution of deprivation. The results are presented using tables, graphs, and maps, forming a basis for discussion in the next chapter.

4.1 MLP prediction accuracy

To evaluate the accuracy of my model, a quantile-binning approach was used to compare the predicted and true IMD scores. The true and predicted IMD scores were divided into a number of equal-frequency bins using pd.qcut() [15], meaning that all bins contain the same quantity of values. The accuracy was calculated by measuring the proportion of instances where the predicted bin matched the true bin. The results over three runs are presented in table 2.

```
true_values_bins = pd.qcut(y_test_original.flatten(), 3, labels=False)
pred_values_bins = pd.qcut(y_pred_original.flatten(), 3, labels=False)

print("True Val:", true_values_bins[:10])
print("Pred:", pred_values_bins[:10])
```

```
True Val: [2 0 0 2 2 0 0 2 0 1]
Pred: [2 0 0 2 2 0 0 2 0 1]
```

Fig. 15. Actual and predicted scores divided into 3 equal-frequency bins

	Run 1	Run 2	Run 3	Median
3	0.838	0.8317	0.8408	0.838
4	0.7604	0.769	0.7646	0.7646
5	0.7027	0.7026	0.7005	0.7026

Table 2. Number of equal frequency bins against the accuracy of predictions

4.2 SHAP visualization

As explained in the previous chapter, the SHAP Beeswarm plot ranks features based on importance from top to bottom. The colour gradient reflects feature values, where red represents high values and blue represents low values. Figure 16 shows the summary plot for 1000 samples. A positive SHAP value means the feature increases the predicted IMD score, while a negative SHAP value decreases it. For example, the plot shows that relatively high values of the 'level 2 or below feature' have a positive impact on the model's predictions.

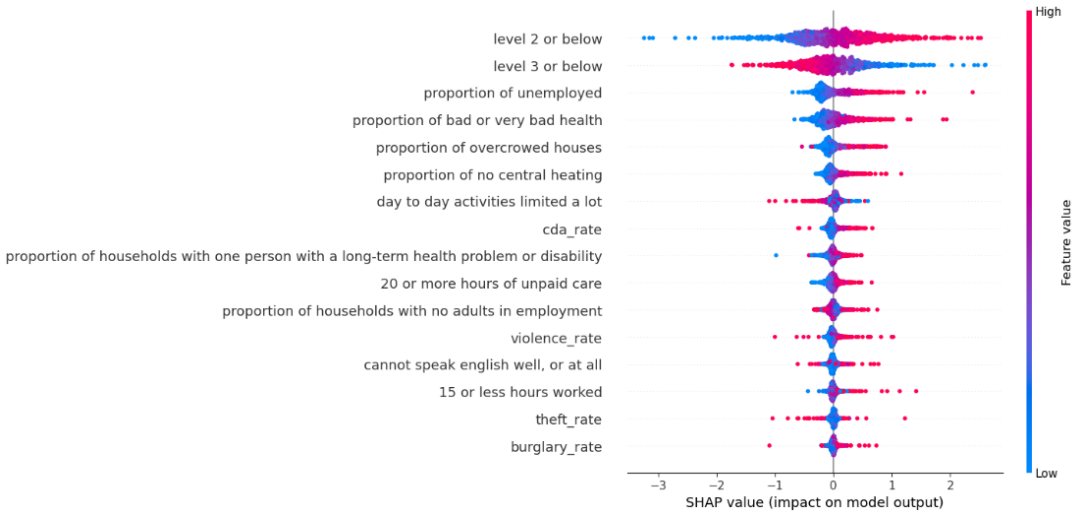

Fig. 16. SHAP summary plot for 1000 samples

4.3 Folium choropleth maps

The following maps illustrate the geographical distribution of deprivation. The colour gradient represents the magnitude of the IMD score. Darker colours (red) indicates higher deprivation, lighter colours (yellow) indicate lower deprivation and black indicates missing data.

4.3.1 Actual 2015 IMD scores. This map presents the officially recorded IMD scores from 2015 for each Lower Layer Super Output Area (LSOA) using 2011 boundaries [17]. This allows for the comparison of trends between actual and predicted IMD scores.

Fig. 17. Actual 2015 IMD Scores

4.3.2 Predicted 2011 IMD scores. Using the trained Multi-Layer Perceptron (MLP) model, IMD scores were estimated for 2011 LSOAs based on census and crime data [18]. Although applying a model onto the dataset is trained on is considered bad practise, it serves as a benchmark for comparison between the IMD scores of the two datasets.

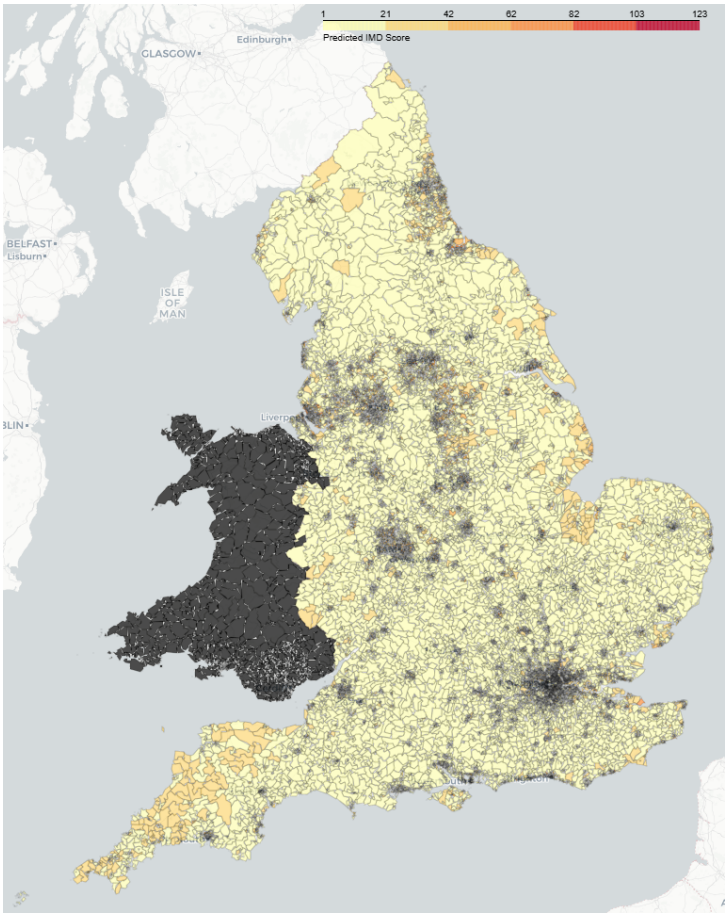

Fig. 18. Predicted 2011 IMD Scores

4.3.3 Predicted 2021 IMD scores. The model was then applied to 2021 census and 2022/23 crime data to predict IMD scores for that period. Since actual 2021 IMD scores were unavailable, this map provides an estimation of how deprivation might have evolved over time [19].

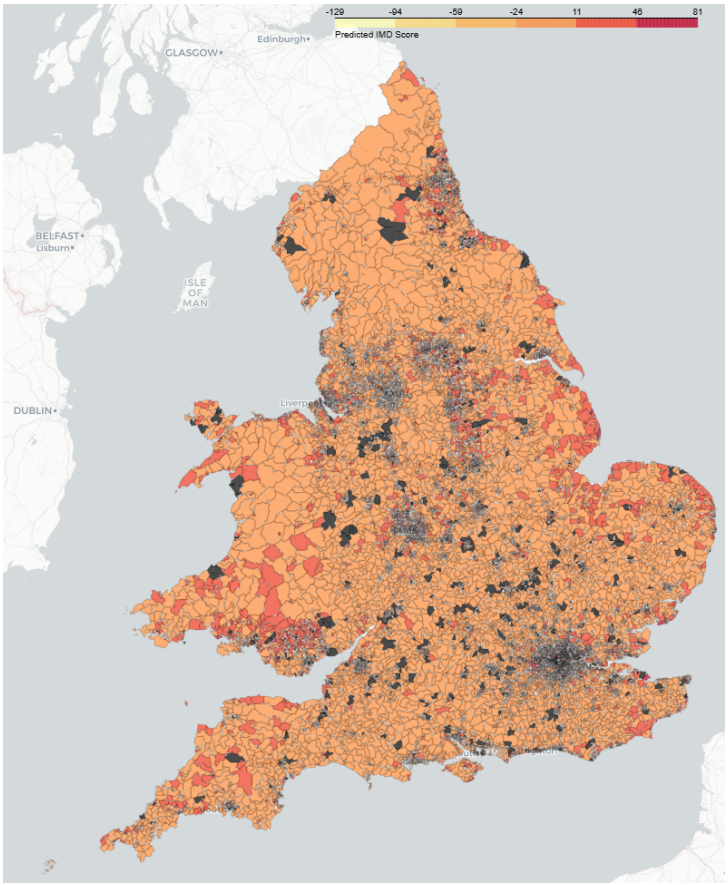

Fig. 19. Predicted 2021 IMD Scores

4.3.4 Changes in IMD scores. This map highlights areas where deprivation has increased or decreased [20]. Red areas signify areas that saw the largest increase in deprivation, while blue areas represent areas with the largest decrease.

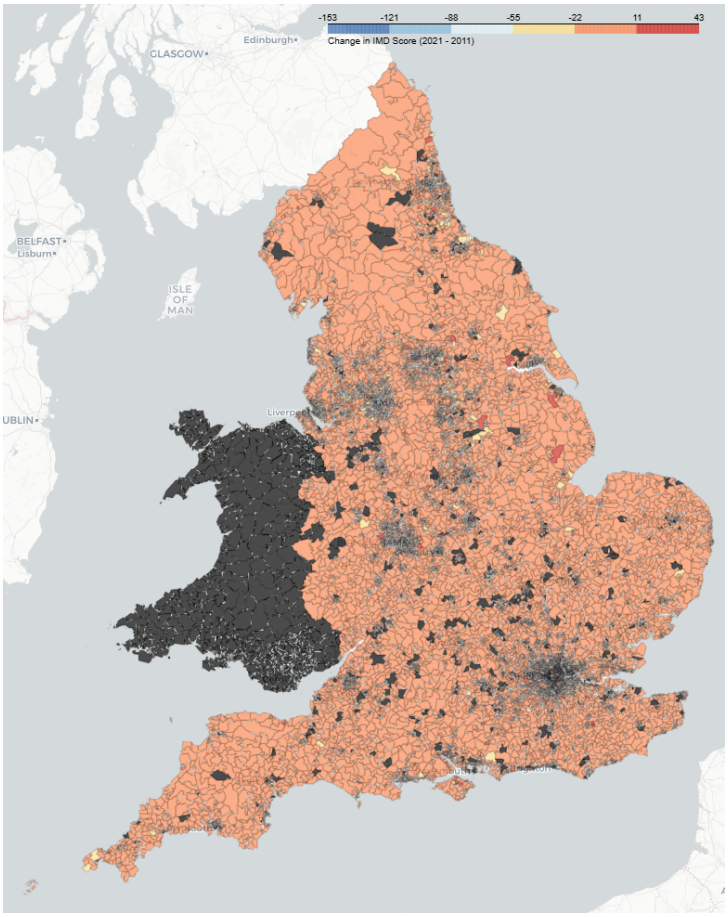

Fig. 20. Changes in IMD Scores (2011–2021)

4.4 Summary

This chapter quantifies the accuracy of the Multi-Layer Perceptron (MLP) model, and presents SHAP visualizations that provide insights into feature importance, highlighting how different variables impact predictions. Finally, four Folium choropleth maps illustrated the geographical distribution of deprivation: actual 2015 IMD scores, predicted 2011 IMD scores, predicted 2021 IMD scores, and the changes in IMD scores from 2011 to 2021. These results and visualisations provide a foundation for further analysis in the next chapter.

5 ANALYSIS

This section presents an evaluation of the MLP model's accuracy, the analysis of SHAP values, and a comparison of predicted and actual IMD scores. Insights will drawn from the SHAP visualisation and Folium maps, in particular, the geographical distribution of changes in IMD scores between the two datasets.

5.1 MLP Accuracy Evaluation

The accuracy seen in table 2 was very consistent through multiple runs which suggests that the model did not suffer from instability or randomness in its predictions. As expected, the highest accuracy is observed when the IMD scores are divided into three bins with a median accuracy of 0.838 compared to 0.7026 with five bins. This shows that the model had a decent ability to capture general deprivation trends, however, there were clear limitations in precise score estimation. This was not unexpected because, whilst the features used in the model were aligned with the domains of the IMD, most did not correspond to the exact indicators used in official IMD calculations. For example, in the case of the income deprivation domain, the indicators used are the receipt of certain income support benefits. This data is not publicly available, so the "Proportion of people working 15 hours or less per week" and "Proportion of unemployed individuals" features were used as a proxy.

5.2 SHAP Beeswarm Plot

This section discusses observations from the SHAP visualisation shown in Figure 16. It is important to note that SHAP values measure direct model impact of features, which may not necessarily reflect their real world importance on the IMD's defenition of deprivation.

5.2.1 Education attainment's Mixed Influence on Deprivation. The feature "level 2 or below" had a positive impact on SHAP values. This means LSOAs with a higher proportion of people with education levels at level 2 or below, which is equivalent to GCSEs at grades A*-C (or 9-4), were predicted to have higher deprivation scores. This aligns with expectations, as lower education levels are linked to deprivation. However, "level 3 or below", equivalent to A-levels, exhibited the opposite trend. Higher values of this feature were associated with lower SHAP values (i.e. lower deprivation). Although unexpected, an explanation of this result is that a larger proportion of individuals with level 3 qualifications may mitigate deprivation to some extent. While the proportion of residents that "cannot speak English well, or at all" matches exactly to one of the indicators used for the education domain in the IMD, it was surprisingly one of the weaker influences. A possible explantation for this is that features like unemployment rate and education level might already be capturing the disadvantages associated with language barriers, and therefore reducing it's contribution to model predictions.

5.2.2 Employment and Income Related Features as Moderate Influences. "Proportion of unemployed residents" had reasonable positive SHAP values, indicating they had a moderate impact

on model predictions. Areas with higher unemployment and poorer health were consistently predicted to have higher deprivation scores. Although the "proportion of households with no adults in employment" and "the proportion of residents that worked 15 hours or less" features followed a similar trend in distribution, they had much less impact on predictions. This may be because the "Proportion of unemployed residents feature" is of the same theme and already captured the impact of income deprivation on the IMD scores, so the other features were made to be less impactful to prevent a double count.

5.2.3 Health and Living Environment features. Of the health related features, "Proportion of residents with bad or very bad health" was the most influential with higher values leading to higher IMD scores (more deprived), with the rest having moderate to low influence. This could again be explained by the reasons given earlier. The "Proportion of overcrowded houses" and the "Proportion of houses with no central heating" features were the exact indicators used for the Living Environment Deprivation Domain of the IMD. Expectedly, these features were relatively influential on model predictions, with higher values producing higher IMD scores.

5.2.4 Crime Related Features' Weak Influence on Predictions. Surprisingly, the crime related features had relatively little influence on predictions, despite being the official indicators used for the crime domain of the IMD. A feasible explanation for this observation is that crime is often correlated to other deprivation indicators. For example, areas with high unemployment, overcrowding, and low education levels tend to have higher crime rates, meaning the model may have already accounted for the impact of crime on the IMD through the other features.

5.3 Analysis of Folium Maps

The maps for actual and predicted IMD scores for both datasets, seen in figures 17, 18 and 19, show the same trends. As shown in figures 21 and 22 deprived areas tend to be in urban or coastal areas, while inland rural areas are typically less deprived. There are many factors as to why urban areas contain a higher proportion of deprived LSOAs. Geographical analysis of the IMD by The Area Based Analysis Unit [2009] found that crime deprivation is concentrated in urban areas, as observed in figure 23. Another obvious factor is higher living costs in urban areas which can lead to higher proportions of overcrowded houses.

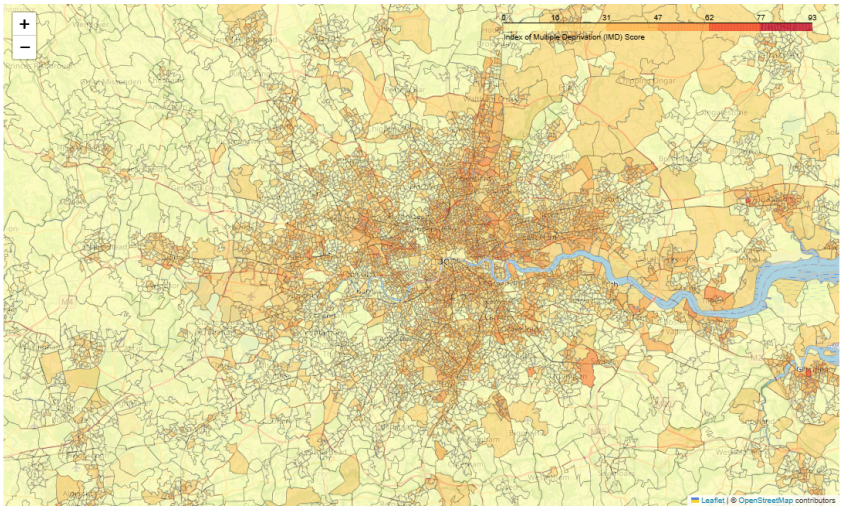

Fig. 21. Zoomed map on London, also showing surrounding towns

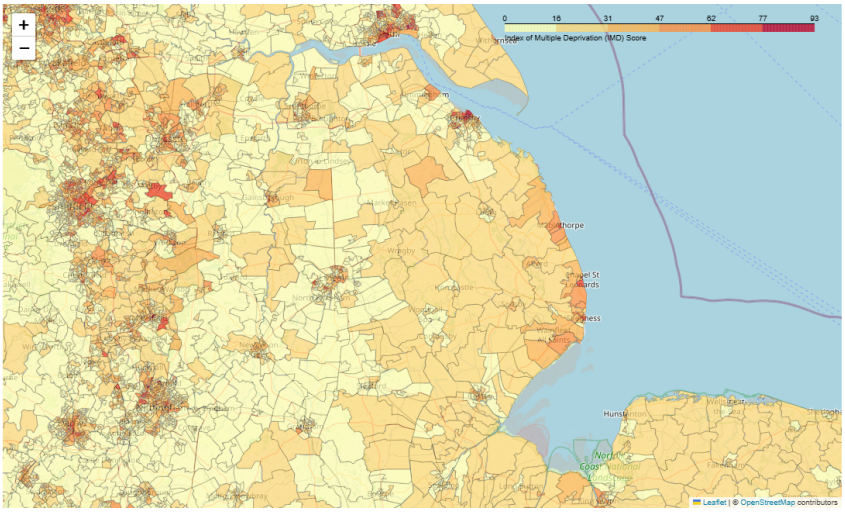

Fig. 22. Zoomed map showing coastal areas in the east of England

Fig. 23. Crime Domain: Most and least deprived LSOAs [The Area Based Analysis Unit 2009]

Although the maps for the predicted IMD for both datasets follow similar trends and contain comparable clusters, there are some clear differences. The mean IMD score saw a significant decrease, (i.e. less deprived). Figure 24 shows summary statistics. Features of the model that saw significant improvements are presented in table 3. Due to potential backwards compatibility issues of some Census statistics, conclusions shouldn't be drawn from IMD score differences between predictions for the two datasets. However, the distribution of change provides interesting insights. Figure 20 shows the distribution of changes in IMD scores. The most important observation is that LSOAs which saw larger decreases in IMD scores were concentrated in urban areas, this can be seen in figure 25. This can be attributed to a lack of investment in rural areas compared to urban areas. In 2021, there was a significant disparity in government spending, with urban areas receiving 44% more public infrastructure funding per person than rural areas [Association 2021].

	IMD predictions using 2011 Census	IMD predictions using 2021 Census	IMD_change
count	31810.000000	31810.000000	31810.000000
mean	21.735498	11.925501	-9.809998
std	14.714138	8.165511	11.703818
min	-0.105391	-45.485012	-97.907147
25%	10.443564	6.647408	-15.410362
50%	17.177425	10.355114	-6.462882
75%	29.562981	15.750901	-1.519240
max	94.011310	73.076797	37.198524

Fig. 24. Summary statistics of the change in predicted IMD scores between the two datasets

Feature	Percentage Decrease of Mean Values
Proportion of overcrowded houses	25.86
Proportion of houses with no central heating	44.63
Proportion of residents with education attainment of level 2 or below	19.32
Proportion of residents who are unemployed	35.36

Table 3. Change in feature values between original and upated datasets

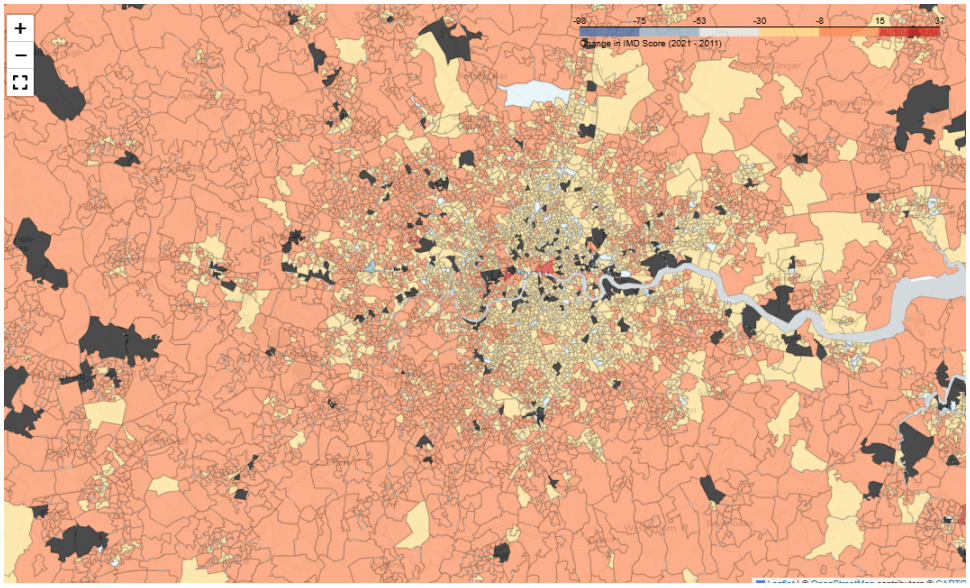

Fig. 25. Change in IMD scores of LSOAs in London and surrounding towns

5.4 Summary

In summary of the analysis of the results, the MLP model achieved stable performance, with the highest accuracy when the IMD scores were divided into three bins (0.838 median accuracy). SHAP analysis showed the varying impacts of the features. Folium maps indicated that urban LSOAs were more deprived but saw better improvements in IMD scores than rural LSOAs, reflecting disparities in government spending.

6 DISCUSSION

This chapter connects the findings of the study to existing research on deprivation. This section will discuss how conclusions from the analysis of my work aligned or contradicted with existing literature. Then, the implications of this studies findings will be explained. Finally, areas for improvement and further research will be discussed.

6.1 Connecting Findings to Existing Research

The SHAP analysis reinforces that education, employment, and health-related factors play significant roles in government defined deprivation. Previous research by McKinney et al. [2012] emphasized the association between educational attainment and deprivation. However, the weaker influence of crime-related features contradicts initial intuition. Though, it could be said that higher crime is produced by higher deprivation in other domains, of which had more impact on model predictions and therefore already captured the impact of crime on deprivation, as discussed in the analysis chapter.

The geographical analysis revealed that, according to the IMDs definition, urban and coastal areas experienced higher levels of deprivation compared to rural areas, but also saw the most significant reductions in IMD scores between the original and updated dataset. This pattern aligns with existing studies (e.g., Bailey and Minton [2017]) that suggest gentrification has played a role, by displacing lower-income populations rather than truly reducing deprivation. Bailey and Minton [2017] conducted an analysis about the "relative concentration of poverty for the 25 largest British cities" between 2004-16. They concluded that while poverty was traditionally concentrated in urban centres, it has increasingly moved to the suburbs. Their explanation of this shift was through various socio-economic factors, including housing affordability, migration patterns, and changes in urban policy which have caused the movement of low-income residents away from urban areas.

6.2 Implications of Findings

The findings of this study have revealed important implications for policymakers seeking to address deprivation. The importance of education attainment to the MLP model aligns with governmental reports [Ministry of Housing, Communities and Local Government 2015] that highlight the role of education in breaking the cycle of poverty. Also, the findings suggesting there is a suburbanisation of poverty, raise concerns about the disparity in government spending between urban and non-urban areas, as well as government handling of the rising costs of living. Rural areas have been receiving less funding per resident than urban areas, whilst the cost of delivering services is higher.

6.3 Limitations and Future Research

A significant limitation of this study was the lack of access to the specific statistics used to calculate the 2015 IMD scores, resulting in suboptimal prediction accuracy even when dividing the scores into a few bins, restricting the conclusions that could be drawn from the results. For further work related to deprivation analysis using machine learning techniques, analysis using the Townsend Deprivation Index [Townsend et al. 1988] instead of the IMD could be more insightful. This measure combines four variables: unemployment, non-car ownership, non-home ownership, and household overcrowding, to measure material deprivation, which are all found in Census data. However, should the statistics used to calculate the IMD be available, it is the better option because it offers a more comprehensive outlook on deprivation.

Another limitation of my work that is recommended for further research is the lack of analysis on the individual domains of the IMD. Training an MLP to predict the scores of each domain would allow for a more detailed analysis about the types of deprivation an area receives and how they have changed when the model is applied to updated data. This could reveal interesting patterns, e.g. maybe coastal towns are more likely to experience one aspect of deprivation than urban areas, but experience relatively less deprivation in other domains. This would reveal not only the areas in need of intervention, but also the specific resources needed which would enable more effective policy making. This method would also allow for the evaluation of how certain policies have impacted areas in specific domains of deprivation.

A final improvement of this work would be to visualise IMD rankings instead of scores. Due to backward compatibility issues with the data, insights gained from comparing IMD scores of the original and updated datasets were limited. However, using rankings, allows for the comparison and better understanding of changes in relative deprivation of areas over time.

7 CONCLUSION

The chapter serves as the final conclusion of this dissertation, summarizing the findings and contributions. It revisits the initial motivations, aims and objectives that inspired this research.

7.1 Motivation and Goals

This study is motivated by the disparity in deprivation across areas in England, which affects access to services and opportunities. Because existing methods for measuring deprivation are resource-intensive, this research explores using machine learning to improve efficiency and complement traditional approaches, providing a quicker and scalable tool for policy development to improve deprivation levels.

7.2 Contributions

Although the accuracy of my model was not ideal, the core objectives of this study were achieved. A neural network was developed, evaluated and applied to updated statistics. A comparative analysis was conducted, the most influential features were examined, and valuable insights were gained. This work, alongside the recommendations for future research in section 6.3, demonstrates the potential use cases for applying machine learning techniques in deprivation analysis.

REFERENCES

A. Abascal, N. Rothwell, A. Shonowo, D. R. Thomson, P. Elias, H. Elsey, G. Yeboah, and M. Kuffer. 2022. Domains of deprivation framework for mapping slums, informal settlements, and other deprived areas in LMICs to improve urban planning and policy: A scoping review. *Health and Place* (2022). https://www.sciencedirect.com/science/article/pii/S019897152200014X [Accessed 17 Nov. 2024].

AIML. 2024. What is a Multilayer Perceptron (MLP)? https://aiml.com/what-is-a-multilayer-perceptron-mlp/ [Accessed 21 Nov. 2024].

The English Rural Housing Association. 2021. Towards a Greener Green Book Process. https://englishrural.org.uk/wp-content/uploads/2021/02/Towards-a-greener-Green-Book-Process.pdf Accessed: 2025-03-27.

Nick Bailey and Jane Minton. 2017. The suburbanisation of poverty in British cities, 2004-16: extent, processes and nature. *Urban Geography* 39, 6 (2017), 892–915. https://doi.org/10.1080/02723638.2017.1405689

F. De Fausti, G. Corbellini, and P. Pellegrini. 2022. Multilayer Perceptron Models for the Estimation of the Attained Level of Education in the Italian Permanent Census. *Statistical Journal of the IAOS* 38, 4 (2022), 637–646. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji210877 [Accessed 19 Nov. 2024].

I. Deas, B. Robson, C. Wong, and M. Bradford. 2003. Measuring Neighbourhood Deprivation: A Critique of the Index of Multiple Deprivation. *Environment and Planning C: Government and Policy* 21, 6 (2003), 883–903. https://doi.org/10.1068/c0240

Department for Work and Pensions. 2012. Income Related Benefits: Estimates of Take-up in 2009-10. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/222915/tkup_full_report_0910.pdf [Accessed 17 Nov. 2024].

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *arXiv preprint arXiv:1705.07874* (2017).

L. Mayhew, G. Harper, and A. M. Villegas. 2020. An investigation into the impact of deprivation on demographic inequalities in adults. *Annals of Actuarial Science* 14, 2 (2020), 400–421. https://www.cambridge.org/core/journals/annals-of-actuarial-science/article/an-investigation-into-the-impact-of-deprivation-on-demographic-inequalities-in-adults/01C4674A085E59BEAD97F4EBE070AC86 [Accessed 18 Nov. 2024].

S. McKinney, S. Hall, K. Lowden, M. McClung, and L. Cameron. 2012. The relationship between poverty and deprivation, educational attainment and positive school leaver destinations in Glasgow secondary schools. *Scottish Educational Review* 44, 1 (2012), 33–45. https://doi.org/10.1163/27730840-04401004

Ministry of Housing, Communities and Local Government. 2015. English Indices of Deprivation 2015: Technical Report. https://www.gov.uk/government/publications/english-indices-of-deprivation-2015-technical-report [Accessed 15 Nov. 2024].

Ministry of Housing, Communities and Local Government. 2019. The English Indices of Deprivation 2019: Statistical Release. https://assets.publishing.service.gov.uk/media/5d8e26f6ed915d5570c6cc55/IoD2019_Statistical_Release.pdf [Accessed 19 Nov. 2024].

P. Montebruno, S. Risch, K. Schurer, and N. T. Longford. 2020. Machine learning classification of British entrepreneurs in census data. *Information Processing Management* 57, 6 (2020), 102210. https://doi.org/10.1016/j.ipm.2020.102210

Office for National Statistics (ONS). 2001. Census 2001: Geography. https://assets.publishing.service.gov.uk/media/5a7c7c9ee5274a559005a337/2001-Census_geography.pdf [Accessed 19 Nov. 2024].

Office for National Statistics (ONS). 2011. 2011 Census Data Finder. Nomis. https://www.nomisweb.co.uk/census/2011/data_finder [Accessed 17 Nov. 2024].

Office for National Statistics (ONS). 2021. 2021 Census Data Finder. Nomis. https://www.nomisweb.co.uk/census/2021/data_finder [Accessed 17 Nov. 2024].

Office for National Statistics (ONS). 2024. Census 2021 geographies. https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeographies/census2021geographies [Accessed 19 Nov. 2024].

K. Schurer, E. Higgs, and Findmypast Limited. 2024. Integrated Census Microdata (I-CeM), 1851-1911. [Data collection]. 2nd edn. https://doi.org/10.5255/UKDA-SN-7481-3 [Accessed 19 Nov. 2024].

Social Metrics Commission. 2024. Measuring poverty 2024. https://socialmetricscommission.org.uk/wp-content/uploads/2024/11/SMC-2024-Report-Web-Hi-Res.pdf [Accessed 17 Nov. 2024].

Office for National Statistics The Area Based Analysis Unit. 2009. Understanding patterns of deprivation. *Regional Trends* (2009).

P. Townsend. 1979. *Poverty in the United Kingdom: A survey of household resources and standards of living*. Penguin Books. https://www.poverty.ac.uk/free-resources-books/poverty-united-kingdom [Accessed Nov. 2024].

P. Townsend. 1987. Deprivation. *Journal of Social Policy* 16, 2 (1987), 125–146. https://www.cambridge.org/core/journals/journal-of-social-policy/article/deprivation/071B5D2C0917B508551AC72D941D6054 [Accessed 17 Nov. 2024].

Peter Townsend, Patricia Phillimore, and Alastair Beattie. 1988. *Health and Deprivation: Inequality and the North* (1st ed.). Routledge. https://doi.org/10.4324/9781003368885

A APPENDIX: GITHUB LINK

Below is a link to the code, and datasets used for this project.

https://github.com/nxdal02/An-Artificial-Neural-Network-for-Deprivation-Analysis-in-England/blob/main/README.md

B APPENDIX: PROFESSIONAL, LEGAL, ETHICAL AND SOCIAL ISSUES

B.1 Professional and Legal Issues

This project complies with The British Computer Society (BCS) Code of Conduct. The work developed is designed with the objective of contributing positively to the public. The project will comply with GDPR, ensuring that data is only used for the purpose of deprivation analysis. The project only uses datasets that are publicly available, non-personalized and ethically cleared. All technical work will adhere to intellectual property laws and code quality and maintainability will be prioritized.

B.2 Ethical issues

Evaluation involving human subjects is not involved in this work but a user evaluation of an interface to simulate hypothetical policy changes was planned. However, this requirement was replaced. See section 3.1.5 for justifications.

B.3 Social Issues

The project's aims align with social interests, by identifying deprived communities to help positively influence policy decisions. The system is designed to avoid reinforcing existing inequalities ensures that the results intend fair and effective policy interventions. The methodology, limitations, and intentions of this work are transparent.